基于组反馈融合机制的视频超分辨率模型 *

张庆武 [1]，迟小羽 [2]，朱　鉴 [1]，陈炳丰 [1†]，蔡瑞初 [1]

(1. 广东工业大学 计算机学院，广州 510006; 2. 北京航空航天大学 青岛研究院，山东 青岛 266000)

摘　要：视频超分辨率(video super-resolution，VSR)，其目的是利用多个相邻帧的信息来生成参考帧的高分辨率版本。现有的许多 VSR 工作都集中在如何有效地对齐相邻帧以更好地融合相邻帧信息，而很少在相邻帧信息融合这一重要步骤上进行研究。针对该问题，提出了基于组反馈融合机制的视频超分辨模型(GFFMVSR)。具体来说，在相邻帧对齐后，把对齐视频序列输入第一重时间注意力模块，然后，把序列分成几个小组，各小组依次通过组内融合模块实现初步融合。接着，不同小组的融合结果经过第二重时间注意力模块。然后，各小组逐组输入反馈融合模块，利用反馈机制反馈融合不同组别的信息，最后，把融合结果输出重建。经验证，该模型具有较强的信息融合能力，在客观评价指标和主观视觉效果上都优于现有的模型。

关键词：视频超分辨率；时间注意力；反馈机制；分组融合

中图分类号：TP391.41　　doi: 10.19734/j.issn.1001-3695.2022.03.0112

Video super-resolution model based on group feedback fusion mechanism

Zhang Qingwu[1], Chi Xiaoyu[2], Zhu Jian[1], Chen Bingfeng[1†], Cai Ruichu[1]

(1. *School of Computers*, *Guangdong University of technology*, *Guangzhou* 510006, *China*; 2. *Qingdao Research Institute of Beihang University*, *Shandong Qingdao* 266000, *China*)

Abstract: Video super-resolution (VSR) , which aims to exploit information from multiple adjacent frames to generate a high-resolution version of a reference frame. Many existing VSR works focus on how to effectively align adjacent frames to better fuse adjacent frame information, and little research has been done on the important step of adjacent frame information fusion. To solve this problem, This paper propose a video super-resolution model based on group feedback fusion mechanism (GFFMVSR) . Specifically, after adjacent frames are aligned, the aligned video sequences are fed into the first temporal attention module. Then, the sequence is divided into several groups, and each group achieves preliminary fusion through the intra-group fusion module in turn. Next, the fusion results of different groups go through a second temporal attention module. Then, each group inputs the feedback fusion module group by group, and uses the feedback mechanism to feedback and fuse the information of different groups. Finally, the fusion result output is reconstructed. It has been verified that the model has strong information fusion ability, and is superior to the existing models in both objective evaluation indicators and subjective visual effects.

Key words: video super-resolution; temporal attention; feedback mechanism; group fusion

0　引言

超分辨率(super-resolution, SR)是指将相应的低分辨率(low-resolution, LR)图像重建为高分辨率(high-resolution，HR)图像的过程。根据输入帧的数量，SR 任务可以可分为两类：单图像超分辨率(single-image super-resolution，SISR)和视频超分辨率(video super-resolution，VSR)。本文是关于视频超分辨率(VSR)任务的研究。VSR 在计算机视觉和图像处理研究领域引起了广泛的关注，具有广泛的应用前景。例如，当监控录像被放大以识别人或车牌时，或者当视频被投影到高清晰度显示器上以获得视觉上的愉悦时，就需要它。

近年来，随着深度学习的发展，基于深度学习的超分辨率算法在性能上有了极大的提高。第一个基于深度学习的 SISR 算法是由 Dong 等人[1]提出的 SRCNN。它由三个卷积层组成，通过端到端的方式学习 LR 图像到 HR 图像的非线性映射，并展示了令人印象深刻的潜力。此后，许多深度学习方法被应用到 SISR 领域。例如，Kim 等人[3]受到 VGG[2]的启发而提出的 VDSR，采用更深层次的卷积网络架构。Li 等人[4]提出了一个通过反馈连接使用更多的上下文信息来纠正低级特征学习的网络架 SRFBN。盘等人[5]提出了一个应用残差中的残差(RIR)和结合使用空间、坐标注意力充分提取和复用特征的网络架构 FFAMSR。尽管这些网络实现了最先进的性能，但高计算成本和内存占用限制了它们在移动设备上的应用。为了解决这个问题，一些轻量级网络被提出来，例如 FALSR-A [6]、SMSR [7]。

在 VSR 领域，Huang 等人[8]提出了一种名为 BRCN 的双向循环卷积网络，可以对跨多帧的长期时间信息进行建模，从而提升了 VSR 的质量。Caballero 等人[9]提出了 VESCPN，

收稿日期：2022-03-15；**修回日期：**2022-04-27　**基金项目：**国家重点研发计划项目(2021ZD011150)；国家自然科学基金优秀青年基金资助项目(6212200101)；广东省自然科学基金资助项目(2016A030310342)；广东省科技计划项目(2016A040403078，2017B010110015，2017B010110007)；广州市珠江科技新星(201610010101)；广州市科技计划项目(201604016075，202007040005)；国家自然科学基金委员会面上项目(61976052)；中国高等教育学会实验室研究专项(21SYYB17)

作者简介：张庆武(1995-)，男，广东茂名人，硕士，主要研究方向为深度学习、视频超分辨率；迟小羽(1980-)，男，北京人，高级工程师，硕导，博士，主要研究方向为机器学习、计算机视觉；朱鉴(1982-)，男，湖南邵阳人，副教授，硕导，博士，主要研究方向为机器学习、计算机视觉；陈炳丰(1983-)，男(通信作者)，广东汕头人，博士，主要研究方向为计算机图形学、高性能计算(chenbf@gdut.edu.cn)；蔡瑞初(1983-)，男，浙江温州人，教授，博导，博士，主要研究方向为数据挖掘、高性能计算．

该网络通过端到端的方式联合训练光流估计和时空网络，从而实现了高效的 VSR。Tao 等人[10]提出了 SPMC，该网络通过设计的亚像素运动补偿模块同时实现了运动补偿和上采样。Kim 等人[11]受 3DCNN 固有的时空学习能力启发提出了 3DSRNet，该网络通过堆叠多个 3D 卷积层进行 VSR 并避免了直接的运动对齐。Jo 等人[12]提出的 DUF 利用 3DCNN 来挖掘时空信息，并预测一个动态上采样滤波器[13]进行隐式运动补偿和上采样，从而代替在像素层面进行的光流估计和对齐。Haris 等人[14]提出的 RBPN 通过使用循环编解码模块来利用空间和时间信息。TDAN[15]和 EDVR[16] 把可变形卷积应用于 VSR 领域并提出了一种时间可变形对齐模块，它们利用该模块在特征层面实现运动对齐。

Hupé[17]和 Gilbert[18]等人发现，在人类认知理论中，连接皮层视觉区域的反馈连接可以将反映信号从高阶区传递到低阶区，从而被加以利用。Zamir 等人[19]更是在前人的基础上提出了一个适用于计算机视觉领域的反馈机制网络。近年来，它已被应用到各种视觉任务[4, 20, 21]的网络架构中，并表现出了不错的结果。据笔者调查，反馈机制还没有在 VSR 的研究领域中得到应用。得益于前人的启发[4, 18]，笔者思考：既然反馈机制[19]允许网络携带历史信息来影响新输入信息的学习，那么融合了部分相邻帧信息的结果对其余相邻帧的融合是否同样具有影响？为此，本文提出了一个基于组反馈融合机制的视频超分辨率模型(video super-resolution model based on group feedback fusion mechanism，GFFMVSR)，本文反馈方案的原理是，具有部分相邻帧信息的结果可以促进其余相邻帧信息的更好融合。

其主要贡献点包括：

a) 提出了基于分组和反馈机制思想的视频超分辨率模型，该模型能有效地融合对齐帧中的高层信息，提高了视频重建的能力。

b) 在视频超分领域内引入了组反馈机制，提供了一种新的相邻帧信息融合方法以提高时空信息融合的性能。

c) 在模型内引入了双重时间注意力，时间注意力模块能捕捉隐藏在相邻帧内的重要信息，使得网络能恢复更清晰，细节更丰富的视频帧。

1 方法论

1.1 网络框架

如图 1 所示，基于组反馈融合机制的视频超分辨率模型主要由五个部分组成：特征提取与对齐模块(feature extraction and alignment module，FEAM)，组内融合模块(intra-group fusion module，IGFM)，双重时间注意力模块(dual temporal attention module，DTAM)，反馈融合模块(feedback fusion module，FFM)，重建模块(rebuild module，RM)。图中蓝色箭头表示反馈融合，绿色箭头表示全局残差跳连接。该网络模型的任务是根据输入的 $2N+1$ 帧视频序列重建参考帧的高分辨率版本。把输入视频序列定义为 $\{I^{LR}_{t-N},\cdots,I^{LR}_t,\cdots,I^{LR}_{t+N}\}$，输出的参考帧超分辨率版本定义为 I^{SR}_t，参考帧的真实高分辨率版本定义为 I^{HR}_t，卷积操作定义为 $Conv(s,n)$，反卷积操作定义为 $Deconv(s,n)$，其中 s 是滤波器的大小，n 是滤波器的数量。

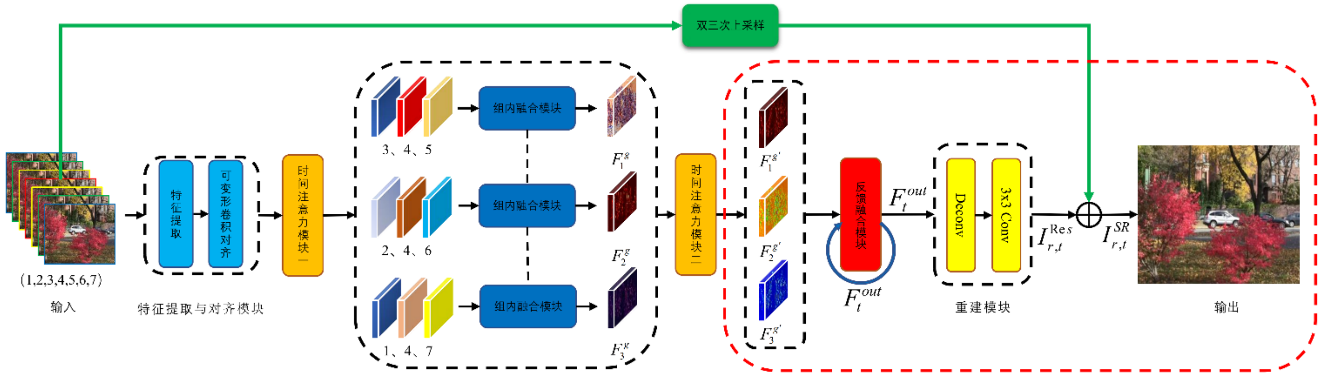


图 1 基于组反馈融合机制的视频超分辨率模型

Fig. 1 Video super-resolution model based on group feedback fusion mechanism

特征提取与对齐模块用于相邻帧特征的提取和对齐，其操作如式(1)所示。

$$\{F^a_{t-N},\cdots,F^a_t,\cdots,F^a_{t+N}\} = f_{FEAM}(\{I^{LR}_{t-N},\cdots,I^{LR}_t,\cdots,I^{LR}_{t+N}\}) \tag{1}$$

其中 $f_{FEAM}(\cdot)$ 代表特征提取与对齐操作。$\{F^a_{t-N},\cdots,F^a_t,\cdots,F^a_{t+N}\}$ 代表经过对齐后的相邻帧特征序列。在 FEAM，特征提取简单地通过具有步进卷积运算的下采样来实现，而对齐操作参考 EDVR[22]中提出的基于多尺度可变形卷积的方法(即 PCD 对齐模块)来实现，该部分建议读者参考 EDVR[22]的 PCD 对齐模块的详细信息。

经过对齐的相邻帧随后输入时间注意力模块一(TAM_1)，从而计算相邻帧与参考帧的相似性，这将有利于组内信息的融合。其操作如式(2)所示。

$$\{F^{a'}_{t-N},\cdots,F^{a'}_t,\cdots,F^{a'}_{t+N}\} = f_{TAM_1}(\{F^a_{t-N},\cdots,F^a_t,\cdots,F^a_{t+N}\}) \tag{2}$$

其中 $f_{TAM_1}(\cdot)$ 代表时间注意力模块一的操作。$\{F^{a'}_{t-N},\cdots,F^{a'}_t,\cdots,F^{a'}_{t+N}\}$ 代表经过时间注意力计算的相邻帧特征序列。

随后对 $\{F^{a'}_{t-N},\cdots,F^{a'}_t,\cdots,F^{a'}_{t+N}\}$ 分成 N 组，每组代表一种特定的帧速率。把各小组序列输入一个参数共享的 IGFM 实现小组内的初步融合，得到融合后的特征序列，定义为 $\{F^g_1,F^g_2,\cdots,F^g_N\}$ (IGFM 模块将在 **1.2 节**中详细阐述)。

融合后的不同组别所蕴涵的信息不一样。为了突出对重建结果有用的信息，在 IGFM 后插入了一个和时间注意力模块一结构相同的时间注意模块二(TAM_2)，构成了双重时间注意力模块(DTAM)，该时间注意力模块将在 **1.3 节**中详细阐述。经过 TAM_2 后的特征序列定义为 $\{F^{g'}_1,F^{g'}_2,\cdots,F^{g'}_N\}$，其操作如式(3)所示。

$$\{F^{g'}_1,F^{g'}_2,\cdots,F^{g'}_N\} = f_{TAM_2}(\{F^g_1,F^g_2,\cdots,F^g_N\}) \tag{3}$$

其中，$f_{TAM_2}(\cdot)$ 代表时间注意力模块二的操作。

跨组别信息通过基于反馈机制的反馈融合模块进一步整合。如图 2 所示，图 1 中红色虚线框可以展开成 T 次迭代($T=N$)，t 代表 1 到 T 中的某一次迭代。为了使 FFM 中的隐藏状态携带输出的概念，联系每次迭代的损失。损失函数将在 1.5 节中详细阐述。把序列 $\{F^{g'}_1,F^{g'}_2,\cdots,F^{g'}_N\}$ 中的元素逐一输入 FFM 模块实现反馈融合。此外 $F^{g'}_1$ 被视为初始隐藏状态 $F^{g'}_0$。

FFM 的第 t 次迭代输入包括第 t 组特征 $F^{g'}_t$ 和来自前一次迭代的隐藏状态 F^{out}_{t-1}。F^{out}_t 代表 FFM 的第 t 次输出。其操作如式(4)所示。

$$F^{out}_t = f_{FFM}(F^{out}_{t-1},F^{g'}_t) \tag{4}$$

其中 $f_{FFM}(\cdot)$ 代表 FFM 的操作，并且反馈的真实过程如图 2 所示。FFM 模块将在 1.4 节中详细阐述。

把反馈融合的结果输入重建模块生成残差图像。如图 1

所示，重建模块使用 $Deconv(k,m)$ 将融合后的 LR 特征放大到 HR 特征，并使用 $Conv(3,c_{out})$ 生成网络的残差图像。重建模块的操作如式(5)所示。

$$I_{r,t}^{Res}=f_{RM}(F_t^{out}) \tag{5}$$

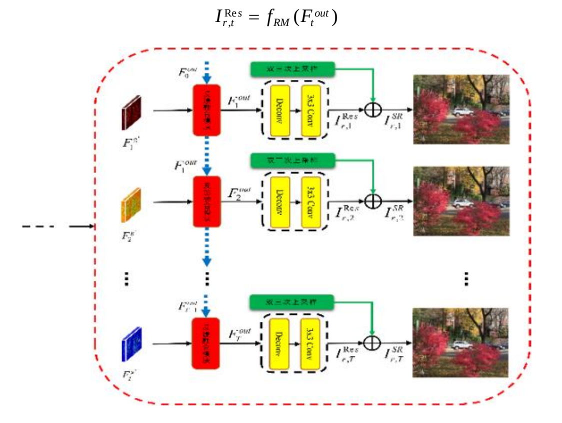


图 2　反馈融合过程展开

Fig. 2　Feedback fusion process unfolds

最后，通过添加网络产生的残差图和输入参考帧的双三次上采样来生成参考帧的高分辨率版本 $I_{r,t}^{SR}$。其操作如式(6)所示。

$$I_{r,t}^{SR}=I_{r,t}^{Res}+f_{up}(I_r^{LR}) \tag{6}$$

其中 $f_{up}(\cdot)$ 代表上采样核的操作。上采样核的选择是任意的，这里使用的是双三次上采样核。在 T 次迭代之后，总共将得到参考帧的 T 个 SR 版本 $(I_{r,1}^{SR},I_{r,2}^{SR},\cdots,I_{r,T}^{SR})$。值得注意的是，随着迭代次数的增加，重建的参考帧携带了越来越多的相邻帧信息，同时也更接近真实的 HR 版本，因此选择最后一次的重建结果作为最终的重建结果。

1.2　组内融合模块(IGFM)

距离较远的相邻帧所隐含的有用信息可能较少。为了充分利用有用信息、剔除过多的无关特征，并提高随后的反馈效率，需要在反馈融合前进行初步的非反馈组内融合。对特征序列 $\{F_{r-N}^{a'},\cdots,F_r^{a'},\cdots,F_{r+N}^{a'}\}$ 进行分组。与之前的工作不一样，基于到参考帧的时间距离，相邻的 $2N$ 帧被分成 N 组。原始序列被重新排列为 $\{G_1,\ldots,G_n\},n\in[1:N]$，其中 $G_n=\{F_{r-n}^{a'},F_r^{a'},F_{r+n}^{a'}\}$ 是由前一帧 $F_{r-n}^{a'}$，参考帧 $F_r^{a'}$ 和后一帧 $F_{r+n}^{a'}$ 组成的子序列，需要提醒的是，参考帧出现在每一组中(具体原因参考第 2.2 节)。不同时间距离的相邻帧的贡献是不相等的，通过分组的方式可以根据参考帧的引导对不同时间距离的相邻帧进行高效的信息提取和融合。值得注意的是，本文的方法可以很容易地推广到任意帧作为输入。

对于每个组，组内融合模块被部署用于每个组内的特征融合。如图 3 所示，该模块的前部分使用具有 卷积核的 3D 卷积层来实现每个小组的时空特征融合。然后，通过在 2D 稠密块中应用 15 个 2D 单元(unit)来深度整合每个组内的信息，最后产生分组特征序列 $\{F_1^g,F_2^g,\cdots,F_N^g\}$。稠密块的每一单元依次由批量归一化[23] (batch normalization，BN)、ReLU[24]、1×1 卷积、BN、ReLU、3×3 卷积组成。如在文献[25]中所做的，每个 2D 单元将所有先前的特征图级联在一起作为输入。最后通过一个 1×1 卷积层减少通道数。2D 稠密块的设计受到 DUF[12]的启发。为了提升效率，组内融合模块的权重由每个组共享，并对本文的数据流通道进行了有效的修改。该模块的操作如式(7)所示。

$$\{F_1^g,F_2^g,\cdots,F_N^g\}=f_{IGFM}(\{G_1,G_2,\cdots,G_N\}) \tag{7}$$

其中 $f_{IGFM}(\cdot)$ 表示卷积操作。 代表组内融合模块操作。第 2.2 节验证了建议的时间分组的有效性。

1.3　双重时间注意力模块(DTAM)

帧间时间关系在 VSR 相邻帧融合中是至关重要的(由于遮挡、模糊区域和视差问题，不同的相邻帧的信息量不同)。时间注意力可以更加聚焦于有利于后续重建的特征上，而非一视同仁。DTAM 由两个一样的，结构如图 4 所示的时间注意力模块构成，分别命名为 TAM_1 和 TAM_2。它们分别聚焦于分组融合前后特征序列时间信息的捕获和权重计算，从而提高信息融合效果。

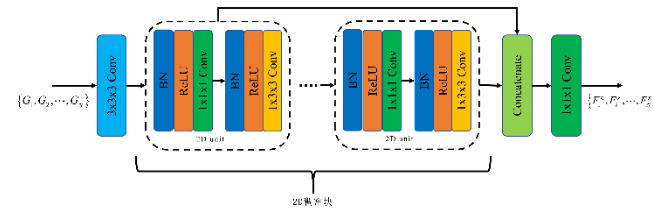


图 3　组内融合模块

Fig. 3　Intra-group fusion module

时间注意力的目标是在一个嵌入空间中计算特征序列的相似性。直观地说，在一个嵌入空间中，应该更多地关注与参考特征更相似的特征信息。在 TAM_1 中，对于每一帧特征，相似性距离 h (即时间注意力图 M_i)可以计算为

$$M_i=h(F_i^a,F_r^a)=sigmoid(\theta(F_i^a)^T\phi(F_r^a)) \tag{8}$$

这里，F_r^a 被视为参考特征。$i\in[r-N,r+N]$。$\theta(F_i^a)$ 和 $\phi(F_r^a)$ 是两个嵌入运算，可通过简单的卷积滤波器来实现。sigmoid 激活函数用于限制输出在[0，1]之间，稳定梯度反向传播。请注意，时间注意力图大小和特征图 M_i 的尺寸是相同的。每帧相邻帧的注意力加权特征计算如下:

$$F_i^{a'}=F_i^a\odot M_i \tag{9}$$

这里，$\odot$ 代表按位置元素的乘法。

同理，对于 TAM_2 也是如此。值得注意的是，在 TAM_2 中参考特征为 F_1^g。其时间注意力图 M_i 和注意力加权特征的计算如式(10)(11)所示(此时 $i\in[1,N]$)：

$$M_i=h(F_i^g,F_1^g)=sigmoid(\theta(F_i^g)^T\phi(F_1^g)) \tag{10}$$

$$F_i^{g'}=F_i^g\odot M_i \tag{11}$$

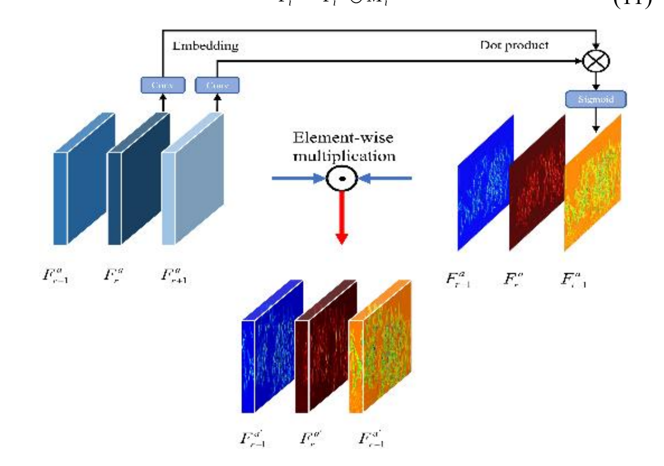


图 4　时间注意力模块

Fig. 4　Temporal attention module

1.4　反馈融合模块(FFM)

FFM 模块如图 5 所示。第 t ($t\in[1:T]$)次迭代的 FFM 接收反馈信息 F_{t-1}^{out} 以指导融合第 t 组特征图 $F_t^{g'}$，然后将融合了更多信息的表示 F_t^{out} 传递给下一次迭代和重构模块，从而形成一个完整的反馈过程。为了实现 FFM 模块的反馈融合功能，该模块依次包含 3 个投影组，其中的信息通过密集的跳跃连接有效地跨层级流动。每个投影组主要包括上采样和下采样操作，该操作可将 HR 特征投影到一个 LR 特征上，从而达到不断细化融合特征的效果。通过迭代执行 FFM 来有效地逐个融合特征序列 $\{F_1^{g'},F_2^{g'},\cdots,F_N^{g'}\}$。迭代过程如图 2 展开所示。

在 FFM 的前端，用 $Conv(1,m)$ 对 $F_t^{g'}$ 和 F_{t-1}^{out} 进行级联和压缩，以通过反馈信息 F_{t-1}^{out} 来指导融合输入特征 $F_t^{g'}$，产生特征细化组的输入特征 LR_t^0：

$$LR_t^0 = C_0([F_{t-1}^{out}, F_t^{g'}]) \tag{12}$$

其中 $C_0(\cdot)$ 代表初始通道压缩操作，$[F_{t-1}^{out}, F_t^{g'}]$ 代表对 F_{t-1}^{out} 和 $F_t^{g'}$ 的级联。定义 H_t^g 和 L_t^g 为第 t 次迭代时 FFM 中第 g ($g \in [1:3]$)个投影组产生的 HR 和 LR 特征图。H_t^g 可以通过以下方式获得：

$$H_t^g = Dec_g([L_t^0, L_t^1, \cdots, L_t^{g-1}]) \tag{13}$$

其中，$Dec_g(\cdot)$ 表示在第 g 个投影组使用 $Deconv_g(k,m)$ 进行上采样操作。相应地，L_t^g 可由下式获得：

$$L_t^g = Conv_g([H_t^1, H_t^2, \cdots, H_t^g]) \tag{14}$$

其中，$Conv_g(\cdot)$ 表示在第 g 个投影组使用 $Conv_g(k,m)$ 进行下采样操作。为降低参数量和提高计算效率，本文在除了第一个投影组外的 $Deconv_g(k,m)$ 和 $Conv_g(k,m)$ 之前添加了 $Conv(1,m)$ 进行通道压缩操作。

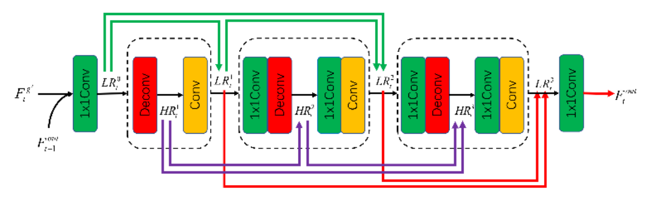


图 5　反馈融合模块

Fig. 5　Feedback fusion module

为了充分利用来自每个投影组的有用信息，本文对投影组产生的 LR 特征进行特征融合(图 5 中的红色箭头所示)，以产生 FFM 模块的输出：

$$F_t^{out} = C_{FF}([L_t^1, L_t^2, L_t^3]) \tag{15}$$

其中，$C_{FF}(\cdot)$ 代表 $Conv(1,m)$ 的函数。

1.5　损失函数

在本工作中，选择 L1 损失来优化所提出的网络。虽然只使用重建序列 $(I_{r,1}^{SR}, I_{r,2}^{SR}, \cdots, I_{r,T}^{SR})$ 中最后一次的结果当做最终结果，但在训练时，仍需要把中间结果与损失函数联系起来，确保每次迭代 FFM 模块都能最大限度融合当前输入特征图的有用信息。网络中的损失函数可以表示为

$$Loss(\theta) = \frac{1}{T}\sum_{t=1}^{T} W^t \| I_r^{HR} - I_{r,t}^{SR} \|_1 \tag{16}$$

其中，θ 表示网络的参数。W^t 是一个常数因子，代表了每次迭代时 SR 结果的贡献值。将所有迭代的 W^t 设置为 1，这意味着每次重建的 SR 结果都有相同的贡献，从而使得每次迭代都能尽可能地去融合高级信息。

2　实验结果和分析

2.1　实验设置

a) 数据集。采用 Vimeo-90k[26]作为训练集，这是一个广泛用于视频超分辨率的训练集。它包含约 90k 份 7 帧的视频剪辑。从高分辨率的视频剪辑中裁剪出空间分辨率为 256×448 的区域。与文献[26，27]相似，通过应用标准差 $\sigma=1.6$ 的高斯模糊核和 4 倍下采样生成 64 × 112 的低分辨率视频剪辑。在两个流行的基准数据集上评估了所提出的方法:Vid4[28]和 Vimeo90K-T[26]。这两个基准数据集都具有各种运动和遮挡的场景，因此适用于评估本文方法的信息融合和高分辨率重建能力。

b) 实现细节：除非另有说明，否则像大多数 VSR 方法[29，16，30]一样，本文网络以 7 个视频帧作为输入，即 $N=3$。使用 PReLU[31]作为每个子网络中除最后一层之外的所有卷积和反卷积层之后的激活函数。将 $Conv(k,m)$ 和 $Deconv(k,m)$ 中的 k 设为 6，以及 4 个步伐和 2 个填充，m 设为 64。使用 Adam[32]优化器进行优化，其中 $\beta_1=0.9$，$\beta_2=0.999$。在训练中不使用权重衰减。学习率最初设置为 2×10^{-4}，然后每 8 个 epoches 降低 0.5 倍，直到 60 个 epoches 结束。小批量的大小设置为 2 训练数据通过 0.5 的概率进行翻转，旋转进行增强。

所有实验都是在配备 Python=3.8、PyTorch=1.1 和 Nvidia 2080TI 的 GPU 服务器上进行。

2.2　消融实验

分组实验。首先用不同的方法来组织输入的序列，一种 Base 方法(记作 Base1)是简单地沿着时间轴堆叠输入序列，并一次性输入 IGFM 和 FFM 模块(中间不具有时间注意力模块),此处的 FFM 模块只执行一次，不具有反馈机制。另外，除了文中建议的分组方式{345，246，147}，还尝试了其他方法的分组:{123，345，567}和{345，142，647}。如表 1 中所示{345，246，147}的分组方法所获得的 PSNR 最高，这暗示了在每组中添加参考帧将有助于模型提取参考帧中缺失的信息。{345，142，647}表现次优则可以归因于距离参考帧不同时间步长的相邻帧信息差异较大，这将不利于信息的分组学习。

表 1　分组方法的消融实验(PSNR)

Tab. 1　Ablation experiment with grouping method

模型	Base1	{123,345,567}	{345,142,647}	{345,246,147}
Vid4	27.09	27.12	27.17	27.20
Vimeo-90K-T	37.06	37.12	37.16	37.19

各模块实验。为了验证各模块的作用，实验中把分组实验提到的{345，246，147}分组方式作为 Base 模型(记作 Base2)，分别在 Base2 模型上引入时间注意力模块一(time attention module 1，TAM_1)，时间注意力模块二(time attention module 2，TAM_2)，反馈融合机制(feedback fusion mechanism，FFM')。值得注意的是，分组后反馈融合机制的关闭是通过在时间维度级联相邻组别特征，然后只执行一次 FFM 模块实现。此外，整合了 TAM_1、TAM_2、FFM' 的完整模型记为 GFFMVSR (Ours)。设置放大的倍数为 4，按照实验设置进行实验后，在 vid4 测试集上的 PSNR 值如表 2 所示。

表 2　在 vid4(4x)测试集上的组件消融实验

Tab. 2　Component ablation experiments on the vid4 (4x) test set

模型	时间注意力模块一	时间注意力模块二	反馈机制	PSNR
Base2				27.20
TAM_1	√			27.29
TAM_2		√		27.28
TAM_1+TAM_2	√	√		27.33
FFM′			√	27.38
GFFMVSR (Ours)	√	√	√	27.47

由表 2 第 1、2、3、4 行可见，引入时间注意力模块一和时间注意力模块二，对 Base2 模型在 PSNR 值上分别有 0.09dB 和 0.08dB 的提升，同时引入两个时间注意力构成双重时间注意力模块后，PSNR 有 0.13dB 的提升。由第 1、5 行可见，引入反馈融合机制，PSNR 值有 0.18dB 的提升。由最后一行可见，整合了双重时间注意力和反馈机制的完整模型性能达到最大值，相比 Base2 模型高出了 0.27dB，这证实了本文提出的模型合理性。

2.3　对比现有先进模型

在本节中，将本文方法与几种最先进的 VSR 方法进行了比较，包括 TOFlow[26]、DUF[12]、RBPN[14]、EDVR[16]、MuCAN[33]、Liu[34]、PFNL[35]和 VSR-Transformer[36]。TOFlow 和 RBPN 都使用光流在像素层面进行显式运动估计。EDVR 则采用对噪声处理能力更强的隐式运动估计。DUF、MuCAN 和 PFNL 则跳过了运动估计过程。最后一种专门使用最新的视觉 transformer (ViT) [37]网络来完成 VSR 任务。通过运行公开的代码或者自己仔细复现了大多数方法，并试图重现原始论文中报告的结果。

Vid4 数据集。表 3 显示了关于 Vid4 的定量结果，其中的数据或者由笔者计算，或者来自于原始论文。其中 Y 和

RGB 分别表示亮度和 RGB 通道，“-”意味着该数值无法取得。作为 GFFMVSR 的降级版本，GFFMVSR-S(只使用时间注意力模块一)在 Y 通道中实现了 27.43/0.8373 的平均 PSNR/SSIM 值，在 RGB 通道中实现了 25.93/0.8186，这可以说是优于所有其他方法。采用双重时间注意力后，GFFMVSR-S 变为 GFFMVSR，在 Y 和 RGB 通道都获得了更高的性能。定性结果如图 6 所示。可以看到 GFFMVSR 比其他方法产生的边缘更锐利，纹理更精细，这也验证了本文方法的优越性。

此外，为了比较时间一致性的性能，从 Vid4 数据集中的日历序列中提取并可视化时间分布图(见图 7)。通过在多个连续的帧中相同位置取水平行的像素(图 7 中的红线)并垂直堆叠它们来获得时间轮廓。可以看出，GFFMVSR 产生了最一致的结果，与其他方法相比，它具有更少的闪烁伪像，并且包含更均匀的线条细节。

表 3　在 Vid4 上的 4×视频超分辨率定量比较

Tab. 3　Quantitative comparison of 4× video super-resolution on Vid4

模型	Frames	Calendar(Y)		City(Y)		Foliage(Y)		Walk(Y)		Average(Y)		Average(RGB)	
	-	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	1	18.83	0.4936	23.84	0.5234	21.52	0.4438	23.01	0.7096	21.80	0.5426	20.37	0.5106
TOFlow[26]	7	22.29	0.7273	26.79	0.7446	25.31	0.7118	29.02	0.8799	25.85	0.7659	24.39	0.7438
DUF-52L[12]	7	24.17	0.8161	28.05	0.8235	26.42	0.7758	30.91	0.9165	27.38	0.8329	25.91	0.8166
RBPN [14]	7	24.02	0.8088	27.83	0.8045	26.21	0.7579	30.62	0.9111	27.17	0.8205	25.65	0.7997
EDVR-L[16]	7	24.05	0.8147	28.00	0.8122	26.34	0.7635	31.02	0.9152	27.35	0.8264	25.83	0.8077
Liu[34]	5	21.61	-	26.29	-	24.99	-	28.06	-	25.23	-	-	-
PFNL[35]	7	24.37	0.8246	28.08	**0.8385**	26.51	0.7768	30.65	0.9135	27.40	0.8384	-	-
VSR-Transformer [36]	5	24.08	0.8125	27.94	0.8107	26.33	0.7635	**31.10**	0.9163	27.36	0.8258	-	-
GFFMVSR-S	7	24.32	0.8253	28.09	0.8312	26.48	0.7771	30.80	0.9158	27.43	0.8373	25.93	0.8186
GFFMVSR	7	**24.39**	**0.8282**	**28.11**	0.8337	**26.54**	**0.7784**	30.85	**0.9166**	**27.47**	**0.8392**	**25.95**	**0.8206**

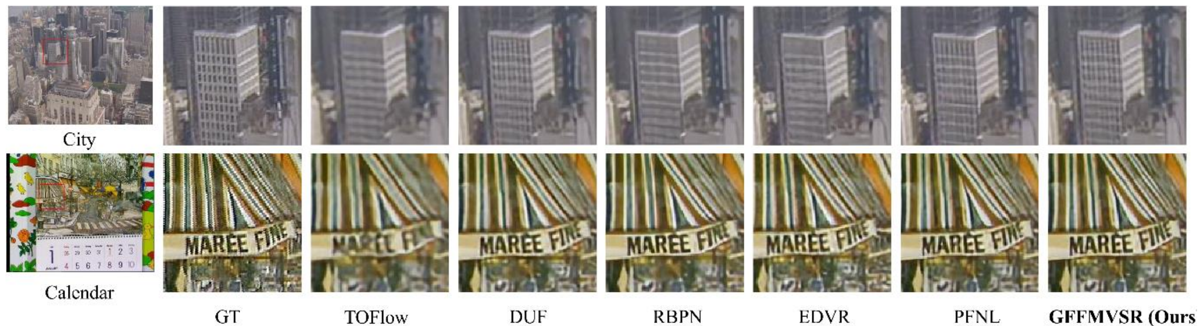


图 6　在 Vid4 数据集上 4×VSR 的定性比较

Fig. 6　Qualitative comparison of 4×VSR on the Vid4 dataset

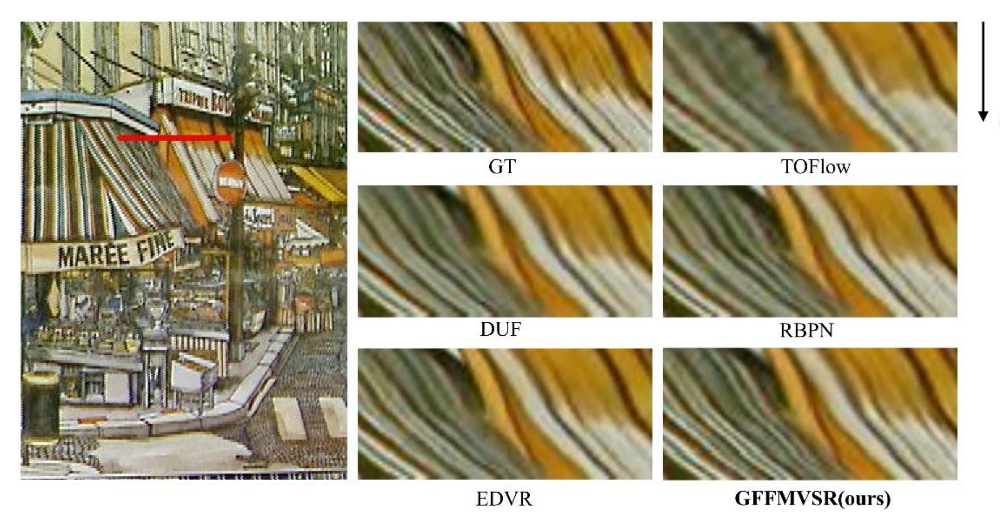


图 7　日历序列上红线的时间轮廓可视化，用于显示时间一致性

Fig. 7　Temporal profile visualization for the red line on the calendar, sequence to show the temporal consistency

Vimeo-90K-T 数据集。Vimeo-90K-T 包含了从 Vimeo-90K 中选取的大约 7K 个视频片段作为测试集，涵盖了大量的场景和大运动。PSNR/SSIM 的定量结果见表 4，其中也包括了大多数方法的参数数量，“-”意味着该数值无法取得。在 PSNR 和 SSIM，本文方法远远超过了大多数最先进的方法，如 TOFlow、DUF、RBPN 和 MuCAN。唯一的例外是 EDVR-L，它的模型大小大约是本文方法的四倍，且 EDVR 涉及到一个需要大量数据和训练时间的预训练过程。尽管如此，本文方法在 PSNR 还是相当不错的，在 SSIM 略胜一筹。

此外，Vimeo-90K-T 的定性结果如图 8 所示。可以看到，GFFMVSR 也可以在这个具有挑战性的数据集上产生视觉上令人信服的 SR 图像。

表 4　在 Vimeo-90K-T 上的 4×视频超分辨率定量比较

Tab. 4　Quantitative comparison of 4× video super-resolution on Vimeo-90K-T

模型	Frames	Param	Y Channel		RGB Channel	
	-	-	PSNR	SSIM	PSNR	SSIM
Bicubic	1	-	31.30	0.8687	29.77	0.8490
TOFlow[26]	7	1.4M	34.62	0.9212	32.78	0.9040
DUF-52L[12]	7	5.8M	36.87	0.9447	34.96	0.9313
RBPN [14]	7	12.1M	37.20	0.9458	35.39	0.9340
EDVR-L[16]	7	20.6M	**37.61**	0.9489	**35.79**	0.9374
MuCAN[33]	7	-	37.32	0.9465	35.49	0.9344
GFFMVSR-S(ours)	7	4.7M	37.43	0.9481	35.62	0.9373
GFFMVSR(ours)	7	4.95M	37.47	**0.9493**	35.68	**0.9385**

3　结束语

本文针对存在于人类视觉系统中的反馈机制仍未在现有视频超分辨率模型中得到充分应用的问题，提出了一种新的端到端可训练的视频超分辨率网络，称为 GFFMVSR。通过将分组思想和反馈机制结合在一起应用到 VSR 任务中，有效地提高了相邻帧信息的融合效果和目标帧重建质量。输入序列被重组为具有不同帧速率的几组子序列。分组允许以分层方式提取时空信息，之后是组内融合模块对小组特征进行初步融合。而反馈融合机制通过模仿人类的认知学习过程，通过反馈信息高效学习并融合新输入的内容。通过在模型的恰当位置应用时间注意力构成的双重时间注意力模型更进一步促使模型专注于有用信息的融合。在几个基准数据集上的大量实验表明，本文提出的模型在定量和定性两方面都优于现有的 VSR 方法。



图 8　在 Vimeo-90K-T 数据集上 4×VSR 的定性比较

Fig. 8　Qualitative comparison of 4×VSR on the Vimeo-90K-T dataset

参考文献：

[1] Dong Chao, Loy Chen Change, He Kaiming, *et al.* Learning a deep convolutional network for image super-resolution [C]// European Conference on Computer Vision. Springer, Cham, 2014: 184-199.

[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv: 1409. 1556, 2014.

[3] Kim J, Lee J K, Lee K M. Accurate image super-resolution using very deep convolutional networks [C]// Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1646-1654.

[4] Li Zhen, Yang Jinglei, Liu Zheng, *et al.* Feedback network for image super-resolution [C]// Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3867-3876.

[5] 盘展鸿，朱鉴，迟小羽，等. 基于特征融合和注意力机制的图像超分辨率模型 [J]. 计算机应用研究, 2022, 39 (3): 5. (Pan Zhanhong, Zhu Jian, Chi Xiaoyu, *et al.* Image super-resolution model based on feature fusion and attention mechanism [J]. Application Research of Computers. 2022, 39 (3): 5.)

[6] Chu Xiangxiang, Zhang Bo, Ma Hailong, *et al.* Fast, accurate and lightweight super-resolution with neural architecture search [C]// 25th International Conference on Pattern Recognition (ICPR) . IEEE, 2021: 59-64.

[7] Wang Longguang, Dong Xiaoyu, Wang Yingqian. *et al.* Exploring sparsity in image super-resolution for efficient inference [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4917-4926.

[8] Huang Yan, Wang Wei, Wang Liang. Bidirectional recurrent convolutional networks for multi-frame super-resolution [J]. Advances in Neural Information Processing Systems, 2015, 28.

[9] Caballero J, Ledig C, Aitken A, *et al.* Real-time video super-resolution with spatio-temporal networks and motion compensation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4778-4787.

[10] Tao Xin, Gao Hongyun, Liao Renjie, *et al.* Detail-revealing deep video super-resolution [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 4472-4480.

[11] Kim S Y, Lim J, Na T, *et al.* 3dsrnet: Video super-resolution using 3d convolutional neural networks [J]. arXiv preprint arXiv: 1812. 09079, 2018.

[12] Jo Y, Oh S W, Kang J, *et al.* Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3224-3232.

[13] Jia X, De Brabandere B, Tuytelaars T, *et al.* Dynamic filter networks for predicting unobserved views [C]// Proceedings ECCV 2016 workshops. 2016: 1-2.

[14] Haris M, Shakhnarovich G, Ukita N. Recurrent back-projection network for video super-resolution [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3897-3906.

[15] Tian Yapeng, Zhang Yulun, Fu Yun, *et al.* Tdan: Temporally-deformable alignment network for video super-resolution [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3360-3369.

[16] Wang Xintao, Chan K CK, Yu Ke, *et al.* Edvr: Video restoration with enhanced deformable convolutional networks [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019: 0-0.

[17] Hupé J M, James A C, Payne B R, *et al.* Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons [J]. Nature, 1998, 394 (6695): 784-787.

[18] Gilbert C D, Sigman M. Brain states: top-down influences in sensory processing [J]. Neuron, 2007, 54 (5): 677-696.

[19] Zamir A R, Wu T L, Sun L, *et al.* Feedback networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1308-1317.

[20] Carreira J, Agrawal P, Fragkiadaki K, *et al.* Human pose estimation with iterative error feedback [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4733-4742.

[21] Sam D B, Babu R V. Top-down feedback for crowd counting convolutional neural network [C]// Thirty-second AAAI Conference on Artificial Intelligence. 2018.

[22] Wang Xintao, Chan K CK, Yu Ke, *et al.* Edvr: Video restoration with enhanced deformable convolutional networks [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019: 0-0.

[23] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [C]// International Conference on Machine Learning. PMLR, 2015: 448-456.

[24] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks [C]// Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, 2011: 315-323.

[25] Huang Gao, Liu Zhuang, Van Der Maaten L, *et al.* Densely connected convolutional networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700-4708.

[26] Xue Tianfan, Chen Baian, Wu Jiajun, *et al.* Video enhancement with task-oriented flow [J]. International Journal of Computer Vision, 2019, 127 (8): 1106-1125.

[27] Jo Y, Oh S W, Kang J, *et al.* Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3224-3232.

[28] Liu Ce, Sun Deqing. On Bayesian adaptive video super resolution [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36 (2): 346-360.

[29] Haris M, Shakhnarovich G, Ukita N. Recurrent back-projection network for video super-resolution [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3897-3906.

[30] Yi Peng, Wang Zhongyuan, Jiang Kui, *et al.* Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3106-3115.

[31] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Delving deep into rectifiers: Surpassing human-level performance on imagenet classification [C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1026-1034.

[32] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint arXiv: 1412. 6980, 2014.

[33] Li Wenbo, Tao Xin, Guo Taian, *et al.* Mucan: Multi-correspondence aggregation network for video super-resolution [C]// European Conference on Computer Vision. Springer, Cham, 2020: 335-351.

[34] Liu Ding, Wang Zhaowen, Fan Yuchen, *et al.* Learning temporal dynamics for video super-resolution: A deep learning approach [J]. IEEE Transactions on Image Processing, 2018, 27 (7): 3432-3445.

[35] Yi Peng, Wang Zhongyuan, Jiang Kui, *et al.* Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3106-3115.

[36] Cao Jiezhang, Li Yawei, Zhang Kai, *et al.* Video super-resolution transformer [J]. arXiv preprint arXiv: 2106. 06847, 2021.

[37] Dosovitskiy A, Beyer L, Kolesnikov A, *et al.* An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv: 2010. 11929, 2020.